Université Paris Cité
**UFR de Mathématiques et Informatique**

# Classification Master 1 IMB
# Examen final du 5 mai 2022 - durée 2h

**On écrira tous les codes R dans un document `Nom_Prenom.R` ou `Nom_Prenom.Rmd`. Les packages suivants seront utilisés : `ade4`, `DALEX`. Utiliser la commande `library` pour charger ces packages. Si un ou plusieurs de ces packages ne sont pas disponibles, utiliser la commande `install.packages` pour les installer.**

## Exercice 1

Soit le tableau de données suivant (5 observations et 4 variables binaires)

| | $X^1$ | $X^2$ | $X^3$ | $X^4$ |
|---|---|---|---|---|
| $e_1$ | 1 | 1 | 0 | 1 |
| $e_2$ | 0 | 0 | 1 | 0 |
| $e_3$ | 1 | 0 | 1 | 1 |
| $e_4$ | 0 | 0 | 1 | 1 |
| $e_5$ | 1 | 0 | 0 | 0 |

1. Entrer le tableau de données en R et appliquer la fonction dist.binary de la librairie `ade4` pour calculer les dissimilarités de Jaccard entre les individus.
2. Quelle est la formule de la dissimilarité de Jaccard telle que calculée dans la fonction dist.binary ? (regarder l'aide de cette fonction).
3. Recalculer sans utiliser R la dissimilarité de Jaccard entre $e_1$ et $e_2$ afin de retrouver la valeur calculée avec R (expliquer le calcul).
4. Effectuer sans utiliser R la classification ascendante hiérarchique obtenue à partir des dissimilarités de Jaccard et en utilisant la stratégie du maximum, en détaillant les étapes. Tracer le dendrogramme obtenu.
5. Vérifier les calculs précédents en effectuant la classification hiérarchique à l'aide de R et afficher le dendrogramme. Retrouve-t-on nécessairement exactement les mêmes étapes et le même dendrogramme en procédant à la main ou avec R ? Si non, les différences ont-elles une importance ?
6. Quel nombre de classes retient-on si l'on suit le critère du saut maximal ? Quelle partition cela donne-t-il ?

Dans la suite, on suppose que l'on dispose en plus d'une autre variable $Y$ binaire pour les 5 observations, telle que $Y_1 = Y_5 = 1$ et $Y_2 = Y_3 = Y_4 = 0$. On cherche à présent à construire un classifieur pour estimer $Y$ à partir des $X^i$.

7. On considère un nouvel individu $e_6$ tel que $X_6^1 = X_6^2 = 1$ et $X_6^3 = X_6^4 = 0$. A l'aide de R, calculer les dissimilarités de $e_6$ avec les autres observations.
8. Déterminer sans utiliser R la classe estimée de $e_6$ suivant la méthode des $k$-plus proches voisins, successivement pour $k = 1$, $k = 3$ et $k = 5$. (N.B. s'il y a des cas d'égalité on affectera aléatoirement).
9. Toujours sans utiliser R, déterminer, successivement pour $k = 1$, $k = 3$ et $k = 5$, la classe estimée suivant la méthode des $k$-plus proches voisins, cette fois pour les individus $e_1$ à $e_5$, et calculer l'erreur apparente des trois classifieurs.
10. Aurait-on pu utiliser la fonction `knn` de R pour utiliser la méthode des $K$-plus proches voisins sur ces données ? Pourquoi ?

**Exercice 2**

Dans cet exercice on va utiliser des données sur les passagers du paquebot Titanic, ayant coulé en 1912.

1. Charger les données `titanic_imputed` du package `DALEX`. Consulter la documentation sur ces données, puis répondez aux questions suivantes :
   - Que contiennent ces données ?
   - Quelle est la différence entre les données sur le Titanic dans ce package `DALEX` et celles du package `stablelearner` ?
   - Quelle est la différences entre les données `titanic_imputed` et les données `titanic` ?
2. Séparer les données en une base d'entraînement contenant 2000 observations et une base de test contenant 207 observations.
3. Utiliser la méthode CART, avec les paramètres par défaut, pour classer les données d'entraînement. Afficher l'arbre de décision obtenu et l'enregistrer dans un fichier PDF (Bouton "Export", puis "Save as PDF")
4. Interpréter l'arbre de décision : quelles remarques "intuitives" sur les facteurs impliquant la survie d'un passager peut-on retrouver en observant l'arbre de décision obtenu ?
5. Calculer le taux d'erreurs obtenu avec l'arbre de décision précédent sur les données de test.
6. Appliquer à présent la méthode des forêts aléatoires sur les données d'entraînement (toujours avec les paramètres par défaut). Calculer le taux d'erreurs obtenu sur les données test et commenter le résultat en le comparant avec celui de la question précédente.

Dans la partie suivante, on va s'intéresser à une méthode permettant de quantifier l'importance d'une variable dans un classifieur, appelée méthode de permutation. L'idée est la suivante : on effectue une permutation aléatoire des valeurs de cette variable sur les données test, en laissant le reste des données de test inchangées. On applique ensuite le classifieur sur ces données modifiées et on calcule le taux d'erreurs $E_{mod}$. Enfin, le score d'importance de la variable est calculé par $E_{mod} - E_{org}$ où $E_{org}$ est le taux d'erreurs original, c'est-à-dire calculé sur les données test sans permutation.

7. Pourquoi cette méthode permet à votre avis de bien quantifier l'importance d'une variable ? Pourquoi ne pourrait-on pas simplement supprimer la variable du tableau de données ?
8. Appliquer la méthode pour le classifieur obtenu précédemment avec la méthode CART, pour la variable "gender" : calculer le tableau de données test modifié, puis le taux d'erreur sur les données test et enfin le score.
9. A présent faire une boucle sur l'ensemble des variables afin de calculer le score d'importance de chaque variable, toujours sur le classifieur CART. Commenter les résultats : d'après ces scores, quelles variables sont déterminantes pour la survie des passagers du Titanic, et quelles variables sont peu importantes ?
10. Comme le score dépend d'une permutation aléatoire, on peut affiner la mesure en faisant une moyenne sur plusieurs tests. Modifier le code de la question précédente pour obtenir des scores moyennés sur 100 permutations. Les conclusions de la questions précédentes sont-elles modifiées ?
11. Enfin effectuer la même analyse en remplaçant le classifieur CART par le classifieur obtenu par la méthode des forêts aléatoires. Commenter les résultats.